DOCUMENT RESUME

ED 456 142 TM 033 204

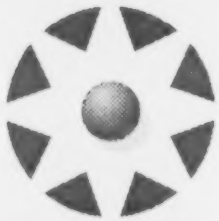
AUTHOR Russell, Michael; Plati, Tom
TITLE Mode of Administration Effects on MCAS Composition Performance for Grades Four, Eight, and Ten. A Report of Findings Submitted to the Massachusetts Department of Education. NBETPP Statements World Wide Web Bulletin.
INSTITUTION National Board on Educational Testing and Public Policy, Chestnut Hill, MA.
SPONS AGENCY Massachusetts State Dept. of Education, Malden.
PUB DATE 2000-05-25
NOTE 53p.
AVAILABLE FROM For full text: http://www.nbetpp.bc.edu.
PUB TYPE Numerical/Quantitative Data (110) -- Opinion Papers (120) -- Reports - Research (143)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Elementary Secondary Education; Essay Tests; Keyboarding (Data Entry); Learning Disabilities; Scoring; Special Education; State Programs; Testing Programs; Writing (Composition)
IDENTIFIERS *Massachusetts Comprehensive Assessment System

ABSTRACT

The studies reported here focus on the effects of mode of administration of the Massachusetts Comprehensive Assessment System (MCAS) in grades 4, 8, and 10. These studies, which draw on earlier studies, also examine the mode of administration effect at different levels of keyboarding speed and for students who received special education (SPED) accommodations in the classroom. Also explored were mode of administration effects for AlphaSmarts and eMates, two portable writing devices. Study participants included 152 4th graders, 228 8th graders, and 145 10th graders who were randomly assigned to write their MCAS essays on paper or on the computer. Essays were scored by teachers and advanced graduate students using MCAS scoring materials, with handwritten essays transcribed and printed on the computer so that scorers would not know whether the essay was originally written by hand or not. Overall, students performed significantly better when they composed essays on the computer. Students who performed both the Language Arts Composition and open-ended questions on the computer scored 4 to 8 points higher on a scale ranging from 200 to 280. The effects were about the same in all grades, but it appeared that students receiving special education services for language arts benefited even more from computer use. (Contains 26 tables, 8 figures, and 56 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

The National Board on Educational Testing and Public Policy



NBETPP

statements

World Wide Web Bulletin · May 25, 2000

ED 456 142

Mode of Administration Effects on MCAS Composition Performance for Grades Four, Eight, and Ten

Michael Russell
National Board on Educational Testing and Public Policy
Center for the Study of Testing, Evaluation and Educational Policy
Boston College, MA
russelmh@bc.edu

Tom Plati
Director of Libraries and Educational Technologies
Wellesley Public Schools
tom_plati@wellesley.mec.edu

May 12, 2000
A Report of Findings Submitted to the
Massachusetts Department of Education

TM033204

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M. Clarke

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Acknowledgements

The studies reported here are a direct outgrowth of Greg Nadeau's recognition of a conflict developing between increased emphasis on computer use in schools and the use of tests as one measure of school accountability. We thank Greg for understanding this issue, having the vision to examine this issue more thoroughly, and for advocating support for these studies within the Department of Education.

We are grateful for the financial support for this study provided by the Department of Education. We are also deeply appreciative of the Brenda Thomas' assistance in providing MCAS materials in a timely manner.

Within Wellesley Public Schools, we thank Matt King, Superintendent of schools, for enabling these studies to occur in his schools. We also thank the principals and teachers who generously contributed their valuable class time to allow their students to participate in this study. In addition, we are indebted to the many teachers who assisted with scoring student responses.

At Boston College, we thank the researchers and graduate students who assisted with data entry, scoring and analyses. We also thank the National Board on Educational Testing and Public Policy for its review and feedback on this report.

Finally, we thank Karen Toomey, our project manager, for the countless hours she invested assuring that computers were operating correctly, tracking down students who had not yet completed questionnaires, proctoring test administrations, assembling databases, and the many more administrative tasks she joyfully completed.

Executive Summary

Both testing and technology have the potential to help improve teaching and learning. Recent research, however, suggests that these two strategies for improving education, namely state level testing programs and writing on computers, may work against each other. Two previous studies provide evidence that students accustomed to writing on computers perform better on written tests when these students are allowed to compose their responses on a computer. In both studies, the magnitude of this improved performance was statistically and practically significant.

Although prior research on computer use and performance on open-ended test items administered on paper does not call into question the value of state level accountability systems, it does suggest that these systems should begin thinking about alternative ways of administering open-ended items. As state level accountability tests begin transitioning from paper administration to computer administration, several issues will arise. First, until all students are accustomed to writing on computers, a better understanding of the extent to which the mode of administration affects student performance at different grade levels must be developed. Second, given the many computing devices available, the affect of performing open-ended items on desktop computers need to be contrasted with performance on cheaper and more portable writing devices such as eMates and AlphaSmarts. Third, before testing programs offer students the option of performing tests on paper or on computer, the extent to which handwritten versus computer printed responses influence raters' scores needs to be explored. Fourth, administrative procedures for administering tests on computers in schools must be developed.

The series of studies presented here focus on the mode of administration effect in grades four, eight and ten. These studies also examine the mode of administration effect at different levels of keyboarding speed and for SPED students. In addition, two studies presented here examine the mode of administration effect for AlphaSmarts and eMates. The extent to which handwritten versus computer printed responses influence raters' scores is also explored. Finally, this series of studies concludes with recommend procedures state level testing programs can employ to provide schools and students the option of performing open-ended items on computers.

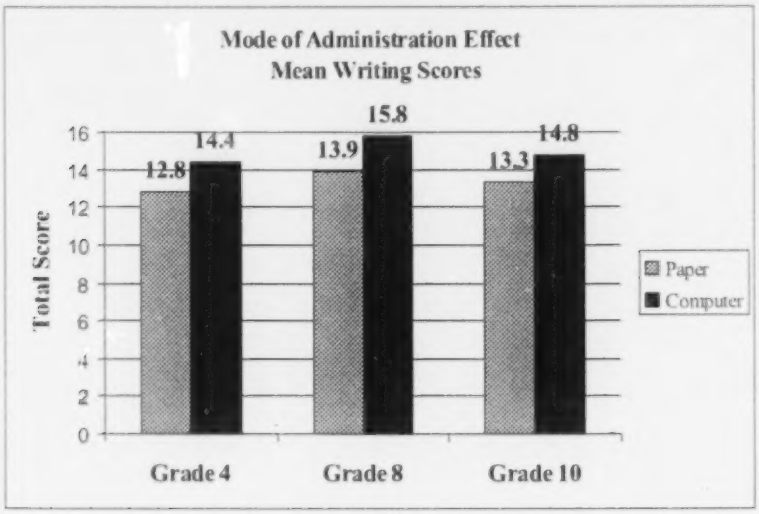
Summary of Studies

Study Details:

- Study occurred during February and March, 2000 in Wellesley (MA) Public Schools
- 152 fourth grade students, 228 eighth grade students, 145 tenth grade students
- Students were randomly assigned to write their essays on paper or on computer
- Same MCAS Language Arts Composition Prompt that was used in the Spring of 1999
- Students had approximately 1 hour to compose a draft and 1 hour to create a final version
- Essays were scored by teachers and advanced graduate students using MCAS Scoring Materials
- All essays composed on paper were transcribed verbatim into the computer and all essays were then printed in the same format so that readers did not know whether the essay was originally written by hand or on a computer

Major Findings:

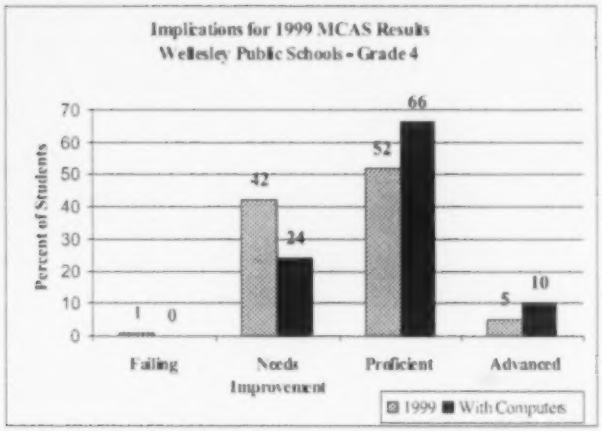
- Students performed significantly better when they composed essays on computer
- Students who performed both the Language Arts Composition and Open-Ended questions on computer would score four to eight points higher on a scale that ranges from 200 to 280
- Effect about the same in grades four, eight and ten
- Students receiving Special Education services for Language Arts may benefit even more by taking written portions of the MCAS Language Arts test on computer

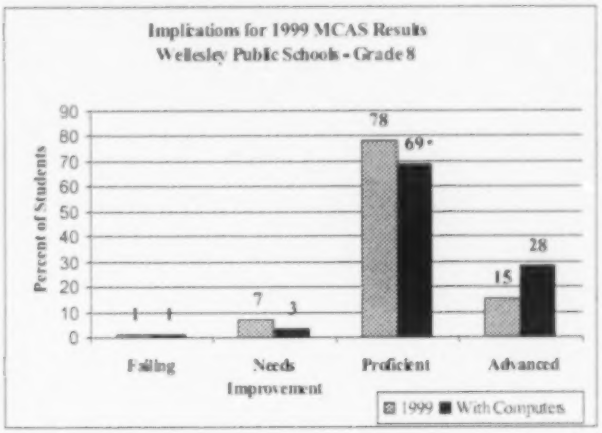


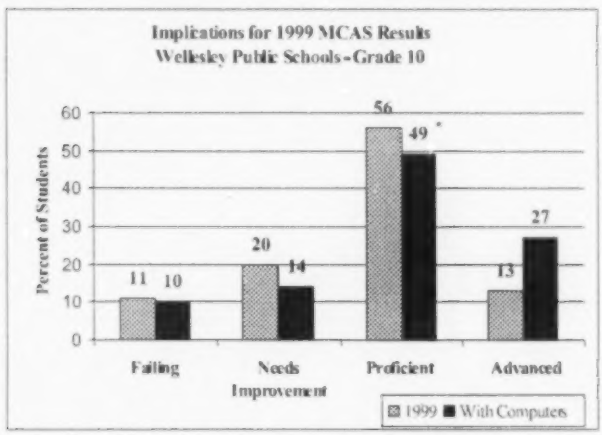
Implications for Wellesley MCAS Language Arts Scores

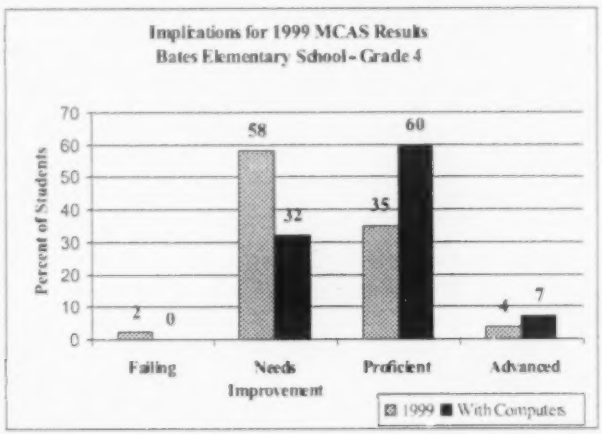
If students could perform both the MCAS Composition item and the four Open-Ended items on computers:

- The percentage of students performing at the "Advanced" Level would double
- The percentage of students performing deemed "Proficient" or better would increase
- The percentage of students "Failing" would decrease









* Note that in both grade 8 and grade 10 a large percentage of students would move from the "Proficient" to the "Advanced" category. A smaller percentage of students move from the "Needs Improvement" to the "Proficient" category. Although it may appear that the percentage of "Proficient" students decreases, the total percentage of students performing at or above the "Proficient" level actually increases.

Implication for State-wide MCAS Language Arts Scores

- Computer use is increasing rapidly across the state.
- Over 75 districts across the state, including Revere, Worcester, Stoughton, North Attleborough and Edgartown, have more high-speed computers per student than does Wellesley.

Sample of District Student to Computer Ratios Across Massachusetts

District	Student to Computer
Acushnet	2.8 to 1
Stoughton	2.9 to 1
Edgartown	3.7 to 1
Taunton	3.9 to 1
Revere	4.3 to1
Ware	4.4 to 1
Worcester	5.2 to 1
North Attleborough	5.2 to 1
Wellesley	5.3 to 1
Springfield	6.2 to1
Boston	8.0 to 1
State Average	7.4 to 1

Note: Over 75 Districts have a better student to computer ratio than Wellesley.

- Within three to four years, most students across the state will use computers regularly for writing
- Based on last year's results, within three to four years, if students could perform both the MCAS Composition item and the four Open-Ended items on computer:
    - The percentage of students performing at the "Advanced" Level would double or triple
    - The percentage of students performing above the "Proficient" Level would increase
    - The percentage of students "Failing" decreases

Recommendation: Provide schools and students the option of performing MCAS open-ended language arts items on paper or on computer.

Related Facts:

- Alberta Learning has provided schools and students the option of performing the Province Graduation English and Social Studies tests on paper or on computer since 1996.
- Recently, Alberta has also given students the option of performing the Biology and French tests on paper or on computer.
- Alberta Learning's Test Administration Manual states:

  "Producing written work on computer is common in contemporary working and learning environments. Students who have been taught to compose on a computer, and who normally produce their written work in this way, would be disadvantaged if they were compelled to hand-write extended assignments."
- In Alberta, the percentage of students opting to perform the English Graduation Test on computer has increased from 9% in 1996 to 30% in 2000.
- In Alberta, students who compose their English Essay on computer consistently perform better than students who compose their response on paper.
- Massachusetts currently has more computers per student than Alberta, Canada.
- Massachusetts does not offer students the option of performing written tests on computer.

Student to Computer Ratios

	Massachusetts		Alberta, CN
	1997	2000	2000
High Speed Computers	15.6 to 1	7.4 to 1	7.7 to 1
Computers of any type	8.4 to 1	5.1 to 1	

Table of Contents

Executive Summary i
Introduction 1
Background 1
Computers and Writing 2
Computers and Testing 3
State Level Testing Programs 5
Study Design 6
Sampling and Group Assignments 7
Prior Computer Use 8
Keyboarding Test 8
Scoring 8
Results 11
Summary Statistics 11
Keyboarding Test 11
Student Questionnaire 11
Indicator of Prior Achievement 13
Composition Scores 14
Comparing Performance by Mode of Administration 15
Grade 8 15
Paper vs. Computer 15
Paper vs. eMate 17
Grade 4 20
Grade 10 22
Examining Keyboarding Speed and Mode of Administration Effect 24
Grade 4 24
Grade 8 25
Grade 10 27
Examining Special Education and Mode of Administration 28
Discussion 30
Limitations 30
Implications 31
Policy Options 34
Administration Guidelines 36
Further Research 38
References 40

Introduction

The prominence of both educational testing and educational technology have increased rapidly over the past decade. When well implemented, state level testing programs encourage schools to think more critically about their curriculum and may provide guidance on topics and skills students need to further develop. When applied to meet curricular goals, education technology provides alternative approaches to sustaining students' interest, developing students' knowledge and skill, and provides supplementary materials that teachers can use to extend student learning. As one example, several studies have shown that writing with a computer can increase the amount of writing students perform, the extent to which students edit their writing (Dauite, 1986, Vacc, 1987; Etchinson, 1989), and, in turn, leads to higher quality writing (Kerchner & Kistinger, 1984; Williamson & Pence, 1989; Hannafin & Dalton, 1987).

Recent research, however, suggests that these two strategies for improving education, namely state level testing programs and writing on computers, may work against each other. As Russell and Haney (2000) describe, two previous studies provide evidence that students accustomed to writing on computers perform better on written tests when these students are allowed to compose their responses using a word processor (without access to spell checker or grammar checker). Despite this improved performance on computer, scores from paper tests are used to make judgments about students, schools, and districts (Sacks, 1999), as well as the impact of technology on student learning.

This paper builds on two previous studies (Russell, 1999; Russell & Haney, 1997) that have explored the effect mode of administration, that is computer versus paper-and-pencil, has on student performance on open-ended items requiring written responses. Whereas the two previous studies have focused on middle school students, the series of studies presented here focus on students in fourth, eighth and tenth grade. In addition, whereas the previous studies examined the effect on relatively short open-ended items that ranged from two to thirty minutes in length, this study focuses on extended composition items designed to be completed during two 45 to 60 minute blocks of time. In addition, two of the three studies presented here introduce a third mode, namely portable writing devices, including eMates and AlphaSmarts.* Similar to laptop computers, these portable writing devices offer a cheaper, but less powerful electronic workspace that may provide an alternative approach to administering tests on desktop computers.

Background

In the past, it was unrealistic to administer tests containing open-ended items on computers in elementary and secondary schools. Until recently, most schools did not have a sufficient number of computers to administer tests in an efficient and timely manner. Moreover, until just a few years ago, most students did not work regularly on computers. The situation, however, is changing rapidly and increasing numbers of students are able to access and write regularly on computers in schools.

While schools had one computer for every 125 students in 1983, they had one for every 9 students in 1995, and 1 for every 6 students in 1998 (Market Data Retrieval, 1999). Similarly, a recent national survey of teachers showed that in 1998, 50 percent of K-12 teachers had students use word processors, 36 percent had them use CD ROMS, and 29 percent had them use the

* eMate is a product of Apple Computers. AlphaSmart is a productive of AlphaSmart, Inc.

WWW (Becker, 1999). Although it is unclear how computers are affecting student achievement in schools (see, for example, Fabos & Young, 1999, questioning the efficacy of Internet based telecommunications exchange programs in schools), there is little doubt that more and more students are writing and performing school assignments on computers.

Computers and Writing

Over the past twenty years, theories about how computers might benefit students' writing have proliferated. To a lesser extent, some researchers have carried out formal studies to examine whether writing on computer actually leads to better writing. Many of these studies have reported that writing on computers leads to measurable increases in students' motivation to write, the quantity of their work and the number of revisions made. Some of these studies also indicate that writing on computers improved the quality of writing. In a meta-analysis of 32 computer writing studies, Bangert-Drowns (1993) reported that about two-thirds of the studies indicated improved quality for text produced on computer. However, the extent to which writing on computers leads to higher quality writing seems to be related to the type of students examined and the extent to which computers are actually used during the writing process. Generally, improvements in the quality of writing produced on a computer are found for learning disabled students, elementary students, low-achieving students and college-aged students. Differences are less frequently found for middle school and high school students, especially when computer use is infrequent.

Learning Disabled, Elementary Students and College-Aged Students

Although neither Kerchner and Kistinger (1984) nor Sitko and Crealock (1986) included a comparison group in their studies, both noted significant increases in motivation, quantity and quality of work produced by learning disabled students when they began writing on the computer. After teaching learning disabled students strategies for revising opinion essays, MacArthur and Graham (1987) reported gains in the number of revisions made on computer and the proportion of those revisions that affected the meaning of the passage. They also noted that essays produced on computer were longer and of higher quality. In a separate study, MacArthur again reported that when writing on a computer, learning disabled students tended to write and revise more (1988). At the first grade level, Phoenix and Hannan (1984) report similar differences in the quality of writing produced on computer.

Williamson and Pence (1989) found that the quality of writing produced by college freshman increased when produced on computer. Also focusing on college age students, Robinson-Stavely and Cooper (1990) report that sentence length and complexity increased when a group of remedial students produced text on the computer. Hass and Hayes (1986) also found that experienced writers produced papers of greater length and quality on computer as compared to those who created them on paper.

Middle and High School Students

In a study of non-learning disabled middle school students, Dauite (1986) reported that although writing performed on the computer was longer and contained fewer mechanical errors, the overall quality of the writing was not better than that generated on paper. In a similar study, Vacc (1987) found that students who worked on the computer spent more time writing, wrote

more and revised more, but that holistic ratings of the quality of their writing did not differ from text produced with paper-and-pencil.

At the middle school level, Grejda (1992) did not find any difference in the quality of text produced on the two mediums. Although Etchison (1989) found that text produced on computer tended to be longer, there was no noticeable difference in quality. Nichols (1996) also found that text produced on computer by sixth graders tended to be longer, but was not any better in quality than text produced on paper. Yet, for a group of eighth grade students, Owston (1991) found that compositions created on computer were rated significantly higher than those produced on paper.

Focusing on high school freshman, Kurth (1987) reports that there was no significant difference in the length of text produced on computer or on paper. Hawisher (1986) and Hawisher and Fortune (1989) also found no significant differences in the quality of writing produced by teenagers on paper and on computer. Hannafin and Dalton (1987) also found that for high achieving students, writing on computer did not lead to better quality writing. But for low-achieving students, texts produced on the computer were of a higher quality than those produced on paper.

Summary of Studies

The research summarized above suggests many ways in which writing on computer may help students produce better work. Although most of this research was performed before large numbers of computers were present in schools, formal studies report that when students write on computer they tend to produce more text and make more revisions. Studies that compare student work produced on computer with work produced on paper find that for some groups of students, writing on computer also has a positive effect on the quality of student writing. This positive effect is strongest for students with learning disabilities, early elementary-aged students and college-aged students. All of the studies described above focus on student work produced in class under un-timed conditions. These studies also focus on work typically produced for English or Language Arts class, such as short stories or essays. However, the series of studies presented here focus on writing produced under formal timed testing conditions. Specifically, this series of studies address the extent to which producing open-ended responses on computer or on paper effects students' performance, particularly for students with different levels of computer use.

Computers and Testing

In 1968, Bert Green, Jr., predicted "the inevitable computer conquest of testing" (Green, 1970, p. 194). Since then, other observers have envisioned a future in which "calibrated measures embedded in a curriculum . . . continuously and unobtrusively estimate dynamic changes in student proficiency" (Bunderson, Inouye & Olsen, 1989, p. 387). Although progress has been made in this area, such visions of computerized testing are far from present reality. Instead, most recent research on computerized testing has focused on computerized adaptive testing, typically employing multiple-choice tests. Perhaps the most widely publicized application of this form of testing occurred in 1993 when the Graduate Record Examination (GRE) was administered nationally in both paper/pencil and computerized adaptive forms.

Naturally, the introduction of computer administered tests has raised concern about the equivalence of scores yielded via computer versus paper-and-pencil administered test versions.

After reviewing several studies that examined the equivalence of scores acquired through computer or paper-and-pencil test forms, Bunderson, Inouye & Olsen concluded that although the mean scores for paper-and-pencil test forms were often slightly higher than for computerized test forms, " [t]he score differences were generally quite small and of little practical significance" (1989, p. 378). Similarly, following their meta-analysis of 29 studies focusing on the equivalence of computerized and paper-and-pencil cognitive ability tests, Mean and Drasgow concluded that their results "provide strong support for the conclusion that there is no medium effect for carefully constructed power tests. Moreover, no effect was found for adaptivity. On the other hand, a substantial medium effect was found for speeded tests" (1993, p. 457).

More recent research, however, shows that young people who have gone to school with computers perform significantly better on open-ended (that is, not multiple choice) questions administered via computer as compared with the same questions administered via paper-and-pencil (Russell, 1999; Russell & Haney, 1997). In both studies, the effect sizes for students accustomed to writing on computer ranged from .57 to 1.25. Effect sizes of this magnitude are unusually large and of sufficient size to be of not just statistical, but also practical significance (Cohen, 1988; Wolf, 1986). Effect sizes of this magnitude, for example, imply that the score for the average student in the experimental group tested on computer exceeds that of 72 to 89 percent of the students in the control group tested via paper and pencil.

Research on the effect of mode of administration on student test performance began with a puzzle. While evaluating the progress of student learning in the Accelerated Learning Laboratory (ALL), a high-tech school in Worcester, MA, teachers were surprised by the results from the second year of assessments. Although students wrote more often after computers were widely used in the school, student scores on writing tests declined in the second year of the new program.

To help solve the puzzle, a randomized experiment was conducted, with one group of sixty-eight students taking math, science and language arts tests, including both multiple-choice and open-ended items, on paper, and another group of forty-six students taking the same tests on computer (but without access to word processing tools, such as spell-checking or grammar-checking). Before scoring, answers written by hand were transcribed so that raters could not distinguish them from those done on computer. There were two major findings. First, the multiple-choice test results did not differ much by mode of administration. Second, the results for the open-ended tests differed significantly by mode of administration. For the ALL School students who were accustomed to writing on the computer, responses written on computer were much better than those written by hand. This finding occurred across all three subjects tested and on both short answer and extended answer items. The effects were so large that when students wrote on paper, only 30 percent performed at a "passing" level; when they wrote on computer, 67 percent "passed" (Russell & Haney, 1997).

Two years later, a more sophisticated study was conducted, this time using open-ended items from the new Massachusetts state test (the Massachusetts Comprehensive Assessment System or MCAS) and the National Assessment of Educational Progress (NAEP) in the areas of language arts, science and math. Again, eighth grade students from two middle schools in Worcester, MA, were randomly assigned to groups. Within each subject area, each group was given the same test items, with one group answering on paper and the other on computer. In

addition, data were collected on students' keyboarding speed and prior computer use. As in the first study, all answers written by hand were transcribed to computer text before scoring.

In the second study, which included about two hundred students, large differences between computer and paper-and-pencil administration were again evident on the language arts tests. For students who could keyboard moderately well (20 words per minute or more), performance on computer was much better than on paper. For these students, the difference between performance on computer and on paper was roughly a half standard deviation. According to test norms, this difference is larger than the amount students' scores typically change between grade 7 and grade 8 on standardized tests (Haney, Madaus, & Lyons, 1993, p. 234). For the MCAS, this difference in performance could easily raise students' scores from the "failing" to the "passing" level (Russell, 1999).

In the second study, however, findings were not consistent across all levels of keyboarding proficiency. As keyboarding speed decreased, the benefit of computer administration became smaller. And at very low levels of keyboarding speed, taking the test on computer diminished students' performance (effect size of about 0.40 standard deviations). Similarly, taking the math test on computer had a negative effect on students' scores. This effect, however, became less pronounced as keyboarding speed increased.

State Level Testing Programs

Despite this improved performance on open-ended items administered on computer, many states are increasingly seeking to hold students, teachers and schools accountable for student learning as measured by state-sponsored tests containing open-ended items administered on paper. According to annual surveys by the Council for Chief State School Officers (CCSSO, 1998), 48 states use statewide tests to assess student performance in different subject areas. Many of these tests are tied to challenging standards for what students should know and be able to do. Scores on these tests are being used to determine whether to: (1) promote students to higher grades; (2) grant high school diplomas; and (3) identify and sanction or reward low- and high-performing schools (Sacks, 1999). Due to the limitations of multiple-choice tests, many statewide tests include sections in which students must write extended answers or written explanations of their work. As the recent CCSSO report commented, "Possibly the greatest changes in the nature of state student assessment programs have taken place in the 1990s as more states have incorporated open-ended and performance exercises into their tests, and moved away from reliance on only multiple-choice tests" (CCSSO, 1998, p. 17).

Although prior research on computer use and performance on open-ended test items administered on paper does not call into question the value of state level accountability systems, it does suggest that these systems should begin thinking about alternative ways of administering open-ended items. As state level accountability tests begin to explore transitioning from paper administration to computer administration, several issues arise. First, until all students are accustomed to writing on computers, a better understanding of the extent to which the mode of administration affects student performance at different grade levels must be developed. Second, given the many computing devices available, the affect of performing open-ended items on desktop computers need to be contrasted with performance on cheaper and more portable writing devices such as eMates and AlphaSmarts. Third, before testing programs offer students the option of performing tests on paper or on computer, the extent to which handwritten versus

computer printed responses influence raters' scores needs to be explored. Fourth, administrative procedures for administering tests on computers in schools must be developed.

The series of studies presented here focus on the mode of administration effect in grades four, eight and ten. In addition, two studies presented here examine the mode of administration effect for AlphaSmarts and eMates. The extent to which handwritten versus computer printed responses influence raters' scores is also explored in the discussion section. This series of studies concludes with procedures state level testing programs can employ to provide schools and students the option of performing open-ended items on computers.

Study Design

To explore these issues, a series of studies was performed in grades four, eight and ten. Students in each of these grade levels responded to an extended composition item from the 1999 Massachusetts Comprehensive Assessment System (MCAS). In Grade 4, one third of the students responded using paper and pencil, one third using a desktop computer, and one third using an AlphaSmart 2000. An AlphaSmart is a portable word processing device that allows students to enter text into a small window that displays four lines of text with forty characters per line. Students may edit text on the AlphaSmart using arrow keys and the delete button. Cutting and pasting was not available to students using an AlphaSmart. To better enable students to revise their composition, students who composed their rough drafts on an AlphaSmart were allowed to edit and finalize their composition on a desktop computer.

In grade 8, the same design was used except the AlphaSmarts were replaced by more powerful and larger screened eMates. An eMate is also a portable word processor, but differs from an AlphaSmart in three important ways. First, the screen is capable of displaying up to twelve lines of text with sixty characters per line. Second, students may use a stylus to select blocks of text and to place the cursor in different locations. Third, in addition to allowing students to cut, copy and paste, eMates also provide a basic spell-checker.

Since students in grade 10 were not familiar with either type of portable device, the study design focused only on paper and pencil versus desktop computer administration. Note that students who worked on paper had access to a dictionary while those who worked on a desktop computer had access to spell-checker.

For all three grade level studies, three types of background information were collected for each student including: prior grades in English, prior computer use, and keyboarding speed.

The study occurred in three stages. During stage 1, prior English grades were collected for each student. For all students, year-end grades from the previous year and mid-term grades from the current year were collected. The course-end grades were used during the stratified group assignment process and the mid-term grades were used as covariates during some analyses.

During stage 2, all students completed the computer use survey and performed the keyboarding test. During stage 3, students performed the composition item.

To the extent possible, the same administration procedures were employed in this study as occurred during the 1999 MCAS administration. In the actual MCAS composition administration, students completed a composition item during two sessions. During the first

session, students composed a first draft. After a fifteen minute break, students then revised and finalized their composition during a second writing session. Both sessions were designed to last for forty-five minutes, but students were given additional time as needed. In some cases, students were reported to take up to an additional hour to complete the composition item. In this study, time constraints and scheduling conflicts challenged efforts to replicate the MCAS composition administration. In grade eight and ten, only two hours of total testing time was available. In grade eight, the two sessions were completed during two consecutive fifty minute blocks. In grades four and ten, however, the two writing sessions were completed over two days.

Sampling and Group Assignment

All students included in this study attended Wellesley Public Schools, a suburban school district located outside of Boston. Within the district, all eighth grade students attending the Middle School participated in the study. In fourth grade, students attending three of the six elementary schools participated. And in the tenth grade, the study included students taking English with any of the four teachers who volunteered their class for this study.

Within each grade level, the process of assigning students to groups differed slightly. In general, students' prior grade in English was used to stratify participating students within each grade level. Participating students within each stratum were then randomly assigned to groups. In grade ten, the composition item was administered in two formats, namely paper and desktop computer. Thus, two groups were formed in grade ten.

In grade four, the composition item was administered in three formats, paper, desktop computer, and AlphaSmart. Students from the three participating schools were pooled and then were randomly assigned to one of three groups.

In grade eight, the composition item was also administered in three formats: paper, computer and eMates. Although an eMate's word processing capabilities are similar to those of a desktop computer, classroom teachers estimated that it would take between four and ten hours of writing for students to become proficient using the stylus to select and move text. In grade eight, approximately ninety students were accustomed to writing with an eMate. Due to time constraints, it was not possible to train other students to use the stylus. For this reason, group assignment occurred in two phases in grade eight. During phase one, students accustomed to working with an eMate were assigned to one sub-group and all other students were assigned to a second sub-group. Within each sub-group, students were stratified by their English grade and then assigned to one of two administration groups. For the eMate sub-group, students were assigned to perform the composition on paper or on eMate. The remaining students were assigned to perform the composition item on paper or on a desktop computer.

Table 1 summarizes the studies conducted within each grade level and indicates the number of students assigned to each group.

Table 1: Summary of Study Designs

	Paper	Computer	AlphaSmart	eMate
Grade 4	49	50	53	--
Grade 8	85	59	--	--

Grade 8	42	--	--	42
Grade 10	74	71	--	--

Prior Computer Use

In addition to performing the composition item, all students completed a computer use questionnaire and a keyboarding test. The computer use questionnaire focused on students' use of computers at home, in school, and during their normal writing process. In addition, the questionnaire collected information about students' use of eMates or AlphaSmarts in school and during the writing process. Finally, the questionnaire queried students about their preference for taking a writing test on: a. paper or on computer, and b. paper or an eMate/AlphaSmart.

Keyboarding Test

To measure keyboarding skills, all students performed a computer based keyboarding test. The keyboarding test contained two passages which students had two minutes apiece to type verbatim into the computer. Words per minute unadjusted for errors were averaged across the two passages and were used to estimate students' keyboarding speed. For the grade eight and ten students, both keyboarding passages were taken directly from encyclopedia articles to assure that the reading level was not too difficult. For the grade four students, the keyboarding passages were taken from a book read in many fourth grade classes.

Although there is considerable debate about how to quantify keyboarding ability (see West, 1968, 1983; Russon & Wanous, 1973; Arnold, et al, 1997; and Robinson, et al, 1979), for the purposes of this study, students average words per minute (WPM) uncorrected for errors was recorded.

Scoring

All responses were scored independently by two raters. Of the fourteen raters employed for this study, nine were full time classroom teachers, four were advanced doctoral students in an educational research program and one was an educational research. All of the raters were blind to the study design, student identities and the mode on which student responses were created. All raters participated in a one-and-a-half to two hour training session prior to scoring student responses.

For all of the items, the scoring criteria developed by MCAS were used. The MCAS scoring guidelines for the composition items focused on two areas of writing, namely Topic/Idea Development and Standard English Conventions. The scale for Topic Development ranged from 1 to 6 and the scale for English Conventions ranged from 1 to 4, with one representing the lowest level of performance for both scales. Table 2 presents the category descriptions for each point on the two scales.

Table 2: Category Descriptions for MCAS Composition Rubrics

Score	Topic Development	English Standards
1	Little topic/idea development, organization, and/or details Little or no awareness of audience and/or task	Errors seriously interfere with communication AND Little control of sentence structure, grammar and usage and mechanics
2	Limited or weak topic/idea development, organization, and/or details Limited awareness of audience and/or task	Errors interfere somewhat with communication and/or Too many errors relative to the length of the essay or complexity of sentence structure, grammar and usage, and mechanics
3	Rudimentary topic/idea development and/or organization Basic supporting details Simplistic language	Errors do not interfere with communication and/or Few errors relative to length of essay or complexity of sentence structure, grammar and usage, and mechanics
4	Moderate topic/idea development and organization Adequate, relevant details Some variety in language	Control of sentence structure, grammar and usage, and mechanics (length and complexity of essay provide opportunity for students to show control of standard English conventions)
5	Full topic/idea development Logical organization Strong details Appropriate use of language	
6	Rich topic/idea development Careful and/or subtle organization Effective/rich use of language	

In addition to the general descriptions, MCAS also provides anchor papers and benchmark papers for each category. These exemplars are grade level specific and respond to the prompt administered at each grade level.

To reduce the influence handwriting has on raters' scores (Powers, Fowles, Farnum & Ramsey, 1994), all responses to the open-ended items administered on paper were transcribed verbatim into computer text. The transcribed responses were randomly intermixed with the computer responses. All student responses were formatted with the same font, font size, line spacing and line width. In this way, the influence mode of response might have on the scoring process was eliminated.

Scoring guidelines designed for each item were used to score student responses. To reduce rater bias all responses were double scored and a spiraled design was employed. At the conclusion of the scoring process, scores awarded by two raters were added together to produce a Topic Development scale that ranged from two to twelve and a English Standards scale that ranged from two to eight.

To estimate inter-rater reliability, the original scores from both raters were used. The resulting scores were compared both via correlation and percent agreement methods. Table 3 shows that for most items the correlation between the two raters' scores ranged from .55 to .64. Agreement within one point ranged from 89% to 100%. Although the inter-rater correlations were lower than desired, they suggest that when discrepancies arose, one set of raters was not consistently more or less lenient than the second set of raters. Although no information has been published regarding inter-rater reliability of composition scores for the actual administration of the MCAS composition items, the extent to which raters were within one point of agreement is similar to the frequency of agreement obtained for the actual MCAS open-ended scoring (Massachusetts Department of Education, 1999).

Table 3: Inter-rater Reliability for Open-Ended Items

	Correlation	% Within 1 Point
Grade 4		
Topic Development	.57	88%
English Standards	.68	96%
Grade 8		
Topic Development	.63	92%
English Standards	.57	100%
Grade 10		
Topic Development	.64	89%
English Standards	.55	97%

Results

This series of studies explores the relationships between prior computer use and performance on an extended composition item administered in grades four, eight and ten. To examine this relationship, three types of analyses were performed within each grade level. First, independent samples t-tests were employed to compare group performance. Second, total group regression analyses were performed to estimate the mode of administration effect controlling for differences in prior achievement. And third, sub-group regression analyses were performed to examine the group effect at different levels of keyboarding speed. However, before the results of these analyses are described, summary statistics are presented.

Summary Statistics

Summary statistics are presented for each grade level included in this study. For the student questionnaire, keyboarding test, and English grades, summary statistics are based on all students included within each grade level. When between group analyses are presented in grade eight, summary statistics for select variables are presented for the sub-set of students that were in the eMate/Paper group or in the Computer/Paper group.

Keyboarding Test

The keyboarding test contained two passages. As described above, the number of words typed for each passage was summed and divided by 4 to yield the number of words typed per minute for each student. Note that due to the passage length, the maximum keyboarding speed students could obtain was 59 words per minute. Table 3 indicates that the mean WPM increased from 24 in grade four to 28 in grade eight to nearly 36 in grade ten.

Table 3: Summary Statistics for the Keyboarding Test

	N	Mean WPM	Std Dev	Min	Max
Grade 4	152	23.71	9.91	5	62
Grade 8	228	27.99	8.89	10	59
Grade 10	145	35.77	10.37	10	59

Student Questionnaire

The student questionnaire contained 12 questions. The maximum score for the Survey was 53 and the minimum score was 2 . The scale for each item varied from 1 to 2 and 0 to 5. To aid in interpreting the summary statistics presented in table 4, the scale for each item is also listed. In addition to mean responses to the individual items, the Survey total score is also presented.

Although comparative data is not available, Table 4 suggests that on average students in all grade levels included in this study have substantial experience working with computers. The average student reports using a computer for three or more years, using a computer in school for one to two hours a week, and using a computer in their home nearly every day. Furthermore, most students report that they use a computer about once a month when writing a first draft. Slightly more students report using a computer to edit the first draft. And most students report using a computer regularly to write the final draft. Similarly, most students indicate that if given the choice, they would prefer to write a paper on computer than on paper.

Table 4 also shows that students in younger grade levels use portable writing devices much more than do students in tenth grade.

Table 4: Summary Statistics for the Student Quesitonnaire

		Grade 4		Grade 8		Grade 10	
	Scale	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Years using computer	1-6	3.6	.76	5.6	.81	5.8	.74
Use computer in school	1-6	1.7	.89	1.2	.70	1.8	1.10
Use computer at home	1-6	2.8	1.25	3.4	1.11	3.4	1.12
Use eMate/AlphaSmart in School	1-6	1.4	.79	.6	.91	0.3	.53
Compose First Draft w/ Computer	1-6	2.2	1.69	3.2	1.74	3.1	1.83
Edit w/ Computer	1-6	2.5	1.50	3.4	1.61	3.4	1.67
Type Final Draft w/ Computer	1-6	3.1	1.38	4.2	1.00	4.1	1.10
Compose First Draft w/ eMate/AlphaSmart	1-6	2.2	1.60	1.0	1.40	0.1	.39
Edit w/ eMate/AlphaSmart	1-6	1.7	1.63	1.0	1.41	0.1	.38
Type Final Draft w/ eMate/AlphaSmart	1-6	1.6	1.52	1.1	1.54	0.1	.38
Paper or Computer Preference	1-2	1.6	.49	1.8	.37	1.7	.47
Paper or eMate/AlphaSmart Preference	1-2	1.4	.50	1.7	.48	1.6	.50
Survey	2-53	25.7	9.05	28.3	7.29	25.6	6.01

Indicator of Prior Achievement

Mid-year English grades were collected for all students included in this study. In fourth grade, students' English grades are composed of four category scores that range from one to four. To calculate student's English grade, the scores from these four categories were summed. The resulting scale ranged from four to sixteen.

For grades eight and ten, alphabetic grades (e.g., A, B-, C+) were awarded. These alphabetic grades were converted to a numeric scale as indicated in Table 5.

Table 5: Letter Grade to Numeric Grade Conversion Chart

Letter Grade	Number Grade
A+	97
A	95
A-	92
B+	87
B	85
B-	82
C+	77
C	75
C-	72
D+	67
D	65
D-	62
F	50

Table 6 displays the mean and standard deviation for mid-year grades for each grade level.

Table 6: Summary Statistics for Mid-Year Grades

	N	Mean	Std Dev	Min	Max
Grade 4	**152**	**11.0**	**1.91**	**8**	**16**
Grade 8	**228**	**87.5**	**5.86**	**65**	**97**
Grade 10	**145**	**82.2**	**6.86**	**62**	**95**

Composition Scores

One extended composition item was administered to students in each grade level. As is explained more fully above, two scores were awarded to each composition. The first score represents the quality of the composition's Topic Development and the second score indicated the quality of the student's Standard English Conventions. Summary statistics are presented in Table 7. In addition to the mean score for students included in this study, the mean score for students across the state who performed the composition in the spring of 1999 are also presented. On average, students included in this study scored higher than students across the state.

Table 7: Summary Statistics for Composition Scores

	N	Scale	Mean	Std Dev	Mean on MCAS
Grade 4					
Topic Dev.	152	2-12	7.64	2.04	6.74
Stand. English	152	2-8	6.02	1.27	5.36
Grade 8					
Topic Dev.	228	2-12	8.36	1.84	7.18
Stand. English	228	2-8	6.29	1.14	5.67
Grade 10					
Topic Dev.	145	2-12	7.60	2.12	7.18
Stand. English	145	2-8	6.45	1.24	5.97

Comparing Performance by Mode of Administration

As is explained in greater detail above, the study designs differed for each grade level. For this reason, results are reported separately for each grade level. Within grade four, an ANOVA was performed to compare mean performance for each mode. In grade eight and grade ten, independent t-tests (assuming equal variances for the two samples and hence using a pooled variance estimate) are employed to examine differences between the pre-assigned modes of administration. The null hypothesis for each of these tests was that the mean performance of the computer and the paper groups and between the eMate and paper groups did not differ from each other. Thus, these analyses test whether performance on computer and on eMate had a statistically significant effect on students' test scores.

To examine whether prior achievement or keyboarding skills differed between the two groups of students who performed each test, independent samples t-tests were also performed for students' mid-term grades and WPM. In addition, independent samples t-tests were performed to examine differences in the length of students' responses.

Grade 8:

A two-step sampling method was used to assign students to groups. First, those students whose teachers allow regular access to eMates were assigned to one group and those students who do not use eMates in English class were assigned to a second group. Students within each group were then randomly assigned to perform the composition item on paper or using an electronic writing device. In effect, this two stage sampling design created two controlled experiments within grade eight. The first experiment contrasts performance on paper with performance on a desktop computer. The second experiment contrasts performance on paper with performance on an eMate. Results for both experiments are presented below.

In addition to comparing student's composition scores and passage lengths, the amount of time students spent working on their compositions was collected for students in grade eight. Given concerns that the MCAS tests consume a large amount of time, testing time was recorded to examine whether drafting and revising on computer might reduce testing time without jeopardizing student's performance. For this reason, testing time is also contrasted.

Paper versus Computer

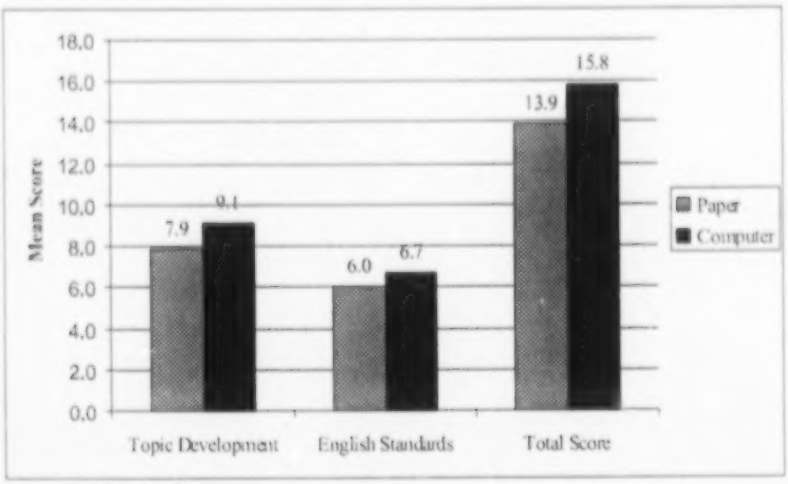
Table 8 displays results for the paper versus computer experiment. Although there was a small difference between the two groups mid-term grades, this difference was not statistically significant. On average, however, students writing on computer produced passages that were one and a half times longer than those composed on paper. More importantly, students who composed on computer also received higher scores for both Topic Development and English Standards. On average, students composing on computer scored 1.2 points higher on Topic Development and .7 points higher on English Standards. When the two sub-scores are combined, the computer group performed nearly two points higher than the paper group (see Figure 1).

Table 8: Between Group Comparisons for Paper versus Computer Experiment

	Mean	Std. Dev.	Std. Error	T-statistic	Significance	Effect Size
Mid-Term Grade						
Paper	86.3	6.11	.67			
Computer	87.4	5.4	.70	1.09	.28	.18
Topic Development						
Paper	7.9	1.8	.20			
Computer	9.1	1.6	.21	4.05	<.001	.64
English Standards						
Paper	6.0	1.2	.13			
Computer	6.7	1.1	.15	3.68	<.001	.61
Total Score						
Paper	13.9	2.8	.30			
Computer	15.8	2.5	.32	4.25	<.001	.69
Passage Length						
Paper	457	144	15.7			
Computer	709	243	31.6	7.77	<.001	1.74
Finish Time						
Paper	107.8	8.2	.89			
Computer	101.9	12.4	1.62	3.45	.001	-.72
WPM						
Paper	26.8	9.8	1.07			
Computer	28.3	7.7	1.01	.97	.36	.15

N for Paper Group=85
N for Computer Group=59

Figure 1: Mode of Administration Effect on MCAS Language Arts Scores – Grade 8



It is also interesting to note that this higher level of performance was produced in less time. Although the difference in Finish Time was statistically significant and represents a sizeable effect size, in reality this six minute difference would have little impact on the total testing time.

Paper versus eMate

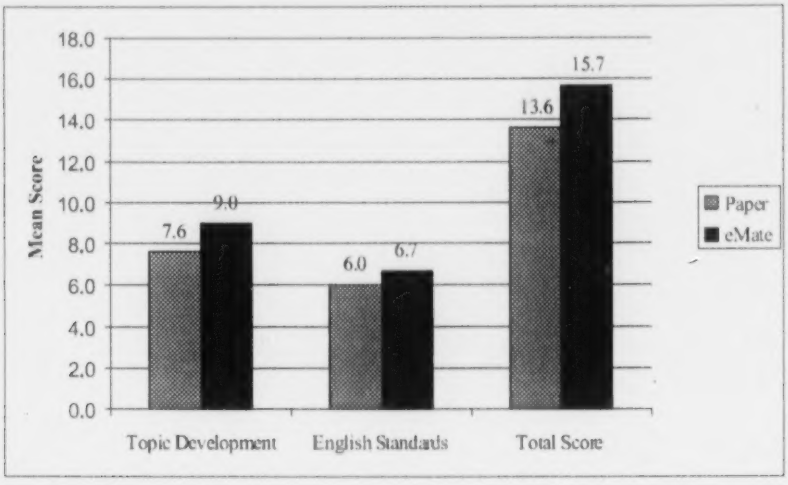
Table 9 displays results for the paper versus eMate experiment. Although the mid-term grades were slightly higher for students in the paper group, this difference was not statistically significant. On average, however, students writing with an eMate produced passages that were twenty percent longer than those composed on paper. Students who composed with an eMate also received higher scores for both Topic Development and English Standards. On average, students composing on eMate scored 1.4 points higher on Topic Development and .7 points higher on English Standards. When the two sub-scores are combined, the eMate group performed over two points higher than the paper group (see Figure 2). In addition, students writing with an eMate finished more than twenty-five minutes faster than did students writing on paper.

Table 9: Between Group Comparisons for Paper versus eMate Experiment

	Mean	Std. Dev.	Std. Error	T-statistic	Significance	Effect Size
Mid-Term Grade						
Paper	89.0	5.7	.87			
eMate	88.5	5.9	.90	.34	.73	-.06
Topic Development						
Paper	7.6	1.5	.23			
eMate	9.0	1.9	.30	3.56	.001	.89
English Standards						
Paper	6.0	.99	.15			
eMate	6.7	.97	.15	3.57	.001	.77
Total Score						
Paper	13.6	2.29	.35			
eMate	15.7	2.70	.42	3.84	<.001	.91
Passage Length						
Paper	448	106	16.4			
eMate	536	160	24.6	2.97	.004	.82
Finish Time						
Paper	107.0	9.4	1.44			
eMate	75.8	9.1	1.41	15.42	<.001	-3.32
WPM						
Paper	29.2	8.2	1.27			
eMate	28.7	9.0	1.40	.28	.78	-.06

N for Paper Group=42
N for eMate Group=42

Figure 2: Mode of Administration Effect on MCAS Language Arts Scores – Grade 8



Note that statistical significance for the t-tests reported above was not adjusted to account for multiple comparisons. Given that seven comparisons were made within each experiment, there is an increased probability that reported differences occurred by chance. Employing the Dunn approach to multiple comparisons (see Glass & Hopkins, 1984), α for c multiple comparisons, α_{pc}, is related to simple α for a single comparison as follows:

$$\alpha_{pc} = 1 - (1-\alpha)^{1/c}$$

Hence, for seven comparisons the adjusted value of a simple 0.05 alpha level becomes 0.007. Analogously, a simple alpha level of 0.01 for a simple comparison becomes 0.001.

Once the level of significance is adjusted for multiple comparisons, the difference in passage length, finish time and all categories of composition scores remain statistically significant for both experiments. Moreover, as shown in Table 8, these differences in composition scores represent effect sizes of .61 to .69 for the computer experiment and effect sizes of .77 to .91 for the eMate experiment (Glass's delta effect size was employed). These effect sizes fall in between those reported by Russell and Haney (1997) and by Russell (1999). For the computer study, these effect sizes suggest that while half of the students in the computer group received total scores above 15.8, approximately 25% of students performing the test on paper scored above 15.8. For the eMate study, the effect size for the total score suggests that while about half of the students writing with an eMate scored above 15.7, less than 19% of students performing the test on paper scored above 15.7.

To control for differences in prior achievement, a multiple regression was performed for each experiment. Tables 10 and 11 present the results of each test score regressed on mid-term grades and group membership. For both regression analyses, the regression coefficient (B) for group membership indicates the effect group membership has on students' performance when the effect of mid-term grade is controlled. Group membership was coded 0 for the paper group and 1 for the computer or eMate group. A positive regression coefficient indicates that performing the test on computer or eMate has a positive effect on students' test performance. A negative

regression coefficient suggests that on average students who performed the test on computer or eMate scored lower than students who performed the test on paper.

Tables 10 and 11 indicate that mid-term grades are a significant predictor of students' scores within both experiments. For each one standard score unit increase in mid-term grades, on average students experience between a .39 and .54 standard score increase in their test score. Tables 10 and 11 also indicate that after controlling for differences in mid-term grades, performing the composition item on either computer or eMate has a positive impact on student scores. This impact ranges from a .25 to .41 standard score increase in student test scores. Although all of these effects are statistically significant, the effects for using eMate are larger than those for using a computer.

Table 10: Composition Scores Regression Analyses for Paper versus Computer Experiment

Topic Development					
R=.61	B	Std. Error	Beta	T	Sig.
Constant	-5.50	1.78			
Mid-Year Grade	0.16	0.02	.52	7.61	<.001
Group	0.96	0.24	.27	3.97	<.001
English Standards					
R=.48	B	Std. Error	Beta	T	Sig.
Constant	-0.50	1.27			
Mid-Year Grade	0.08	0.02	.39	5.17	<.001
Group	0.58	0.17	.25	3.39	.001
Total Score					
R=.60	B	Std. Error	Beta	T	Sig.
Constant	-5.99	2.69			
Mid-Year Grade	0.23	0.03	.51	7.45	<.001
Group	1.55	0.37	.29	4.22	<.001

Table 11: Composition Scores Regression Analyses for Paper versus eMate Experiment

Topic Development					
R=.62	B	Std. Error	Beta	T	Sig.
Constant	-6.7	2.49			
Mid-Year Grade	0.16	0.03	.50	5.75	<.001
Group	1.40	0.31	.39	4.41	<.001
English Standards					
R=.63	B	Std. Error	Beta	T	Sig.
Constant	-2.31	1.14			
Mid-Year Grade	0.09	0.02	.51	5.88	<.001
Group	0.80	0.18	.39	4.54	<.001
Total Score					
R=.66	B	Std. Error	Beta	T	Sig.
Constant	-8.98	3.50			
Mid-Year Grade	0.25	0.04	.54	6.46	<.001
Group	2.20	0.45	.41	4.93	<.001

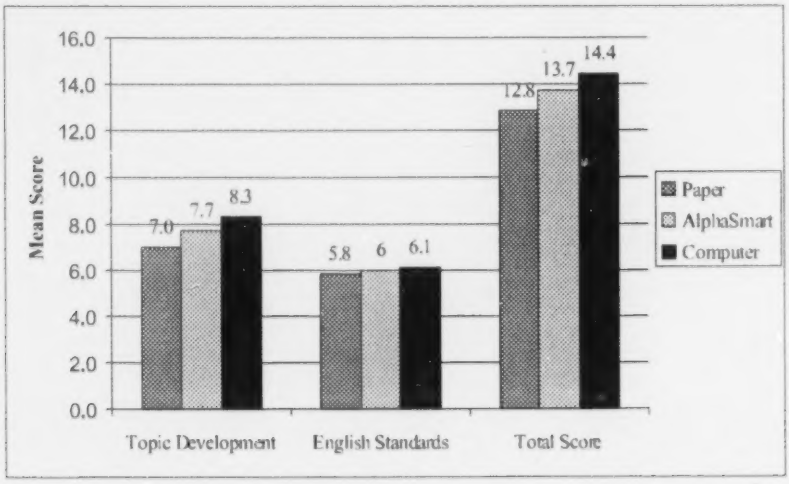
Grade 4:

The study in grade four compared performance on paper with performance on computer and on AlphaSmarts. Students were randomly assigned to one of these three groups. Table 12 indicates that mean scores for Topic Development, Total Score and Passage Length differed among the three groups (see also Figure 3).

Table 12: Summary Statistics by Mode of Administration

	Paper		Computer		AlphaSmart	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Mid-term Grade	10.7	1.71	11.0	1.88	11.2	2.12
Topic Development	7.0	2.01	8.3	1.96	7.7	1.99
English Standards	5.8	1.30	6.1	1.29	6.0	1.22
Total Score	12.8	3.09	14.4	3.12	13.7	2.99
Passage Length	305	141.1	445	258.7	332	159.2
WPM	22.1	9.04	24.2	8.48	24.8	11.76

Figure 3: Mode of Administration Effect on MCAS Language Arts Scores – Grade 4



To examine whether these differences were statistically significant, a one-way analysis of variance was performed. Table 13 indicates that the group means for Topic Development, Total Score and Passage Length did differ significantly among the three groups. Scheffe post-hoc comparisons were then performed for these three variables. Table 14 indicates that none of the

differences between paper and AlphaSmarts were statistically significant. However, the differences between paper and computer were all statistically significant. Specifically, the mean scores for Topic Development and Total Score were significantly higher for students who performed the composition item on computer than was performance on paper. In addition, students who worked on computer wrote longer passages than either students who wrote on paper or on AlphaSmarts.

Table 13: Mode of Administration ANOVA

	Mean Square			
	Between	Within	F	Sig.
Mid-term Grade	3.36	3.67	.92	.40
Topic Development	20.94	3.94	5.31	.006
English Standards	1.18	1.16	.73	.48
Total Score	31.91	9.39	3.40	.03
Passage Length	274,892	37,359	7.36	.001
WPM	99.19	98.26	1.01	.36

Table 14: Scheffe Multiple Comparisons for Topic Development, Total Score and Passage Length

	Paper vs Computer			Paper vs. AlphaSmart			Computer vs. AlphaSmart		
	Mean Diff.	Std. Error	Sig.	Mean Diff.	Std. Error	Sig.	Mean Diff.	Std. Error	Sig.
Topic Development	1.30	.40	.006	.68	.39	.28	.62	.39	.29
Total Score	1.6	.62	.036	.88	.61	.35	.73	.60	.49
Passage Length	140	38.8	.002	27.5	38.3	.77	112	38.1	.014

Although the computer group scored significantly higher than the paper group, the computer group also had slightly higher Mid-term grades. To control for differences in mid-term grades, Topic Development, English Standards and Total Score were each regressed on Mid-term grades and on group membership. For these regressions, two dummy variables were created for group membership. The first, called Computer, was coded as 1 for students who performed the test on computer and 0 for all other students. The second dummy variable, called Alpha, was coded 1 for students who performed the test on an AlphaSmart and 0 for all other students. Table 15 indicates that after controlling for differences in mid-term grades, performing the composition item on computer still had a significant effect on students' Topic Development and Total scores. It is interesting to note that although students working on computer had access to spell-checker, this access did not result in significantly higher English Standards scores. Access to a computer, however, did enable students to write longer passages and, in turn, receive significantly higher scores for Topic Development.

Table 15: Composition Scores Regression Analyses for Grade 4

Topic Development					
R=.54	B	Std. Error	Beta	T	Sig.
Constant	1.53	.83			
Mid-Year Grade	0.51	.07	.48	6.88	<.001
Computer	1.17	.35	.37	3.34	.001
Alpha	0.42	.35	.10	1.21	.23
English Standards					
R=.54	B	Std. Error	Beta	T	Sig.
Constant	2.06	.52			
Mid-Year Grade	0.35	.05	.53	7.67	<.001
Computer	0.21	.22	.08	0.96	.34
Alpha	0.02	.22	.01	.09	.93
Total Score					
R=.57	B	Std. Error	Beta	T	Sig.
Constant	3.58	1.24			
Mid-Year Grade	0.86	.11	.53	7.81	<.001
Computer	11.37	.52	.21	2.64	.009
Alpha	0.44	.52	.07	0.85	.40

Grade 10:

The grade 10 study contrasted performance on paper with performance on a computer. To examine the mode of administration effect, independent samples t-tests were performed for all three composition scores. In addition, the mean mid-term grades, passage length and keyboarding speed were compared.

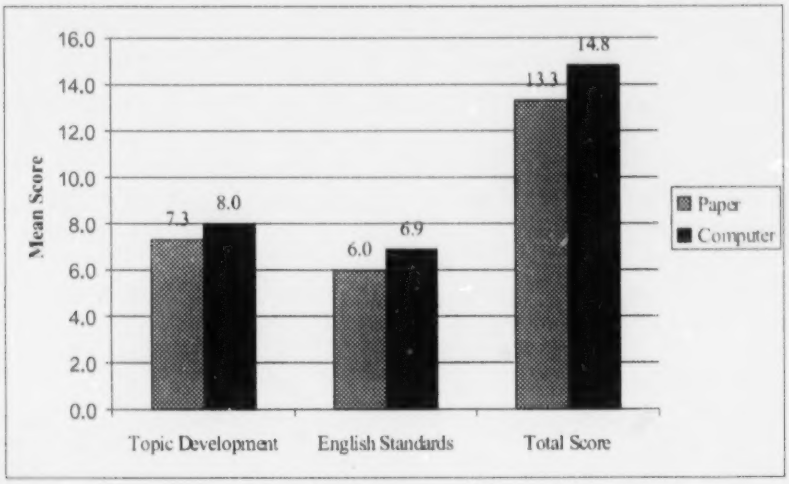
Table 16 indicates that there was a small difference between the two groups mid-term grades. This difference, however, was not statistically significant. On average, passages composed on computer were one hundred words longer than those produced on paper. Responses composed on computer also received higher scores than those produced on paper. The effects were larger for English Standards than for Topic Development. After adjusting significance for multiple comparisons, the difference in both the English Standards and Total score were statistically significant. Moreover, the effect sizes for these differences ranged from .32 for Topic Development, .68 for English Standards and .51 for the Total Score. The effect size for the total score suggests that while about half of the students writing on computer scored above 14.8, approximately 30% of students performing the test on paper scored above 14.8. On average, students who composed on computer scored 1.5 points higher than students who composed on paper (see Figure 4).

Table 16: Between Group Comparisons for Paper versus Computer

	Mean	Std. Dev.	Std. Error	T-statistic	Significance	Effect Size
Mid-Term Grade						
Paper	81.9	6.2	.72			
Computer	82.6	7.5	.89	.59	.56	.11
Topic Development						
Paper	7.3	2.2	.25			
Computer	8.0	2.0	.24	2.02	.05	.32
English Standards						
Paper	6.0	1.3	.15			
Computer	6.9	1.1	.13	4.43	<.001	.68
Total Score						
Paper	13.3	3.1	.36			
Computer	14.8	2.8	.33	3.21	.002	.51
Passage Length						
Paper	462	138	16.1			
Computer	567	292	34.6	2.81	.006	.76
WPM						
Paper	35.1	10.4	1.21			
Computer	36.4	10.4	1.23	.74	.46	.12

N for Paper Group=74
N for Computer Group=71

Figure 4: Mode of Administration Effect on MCAS Language Arts Scores – Grade 10



To control for differences in prior achievement, a series of multiple regressions was performed in which each composition score was regressed on mid-term grades and group membership. Table 17 indicates that mid-term grades are a significant predictor of students'

scores. For each one standard score unit increase in mid-term grades, on average students experience between a .27 and .39 standard score increase in their test score. After controlling for differences in mid-term grades, performing the composition item on computer had a positive impact on student scores. This impact, however, was twice as large for English Standards than for Topic Development. After adjusting for multiple comparisons, the effect was statistically significant for both English Standards and for the Total Score.

Table 17: Composition Scores Regression Analyses for Paper versus Computer Experiment

Topic Development					
R=.42	B	Std. Error	Beta	T	Sig.
Constant	-2.50	1.94			
Mid-Year Grade	0.12	0.02	.39	5.06	<.001
Group	0.62	0.32	.15	1.93	.05
English Standards					
R=.44	B	Std. Error	Beta	T	Sig.
Constant	1.98	1.13			
Mid-Year Grade	.05	0.01	.27	3.63	<.001
Group	.83	0.19	.33	4.44	<.001
Total Score					
R=.46	B	Std. Error	Beta	T	Sig.
Constant	-0.54	2.71			
Mid-Year Grade	0.17	0.03	.38	5.13	<.001
Group	1.45	0.45	.24	3.22	.002

Examining Keyboarding Speed and Mode of Administration Effect

Grade 4

The regression analyses presented above (Table 15) indicate that mode of administration had a significant effect for students who performed the composition item on computer but not for students who worked on AlphaSmarts. To test whether the effect of mode of administration varied for students with different levels of computer skill, students' WPM was used to form two groups. The first group contained students whose WPM was below the mean or less than 24 wpm. The second group contained students whose WPM was above then, or greater than 24 wpm. For each group, the composition total score was regressed on mid-term grades and group membership. As in the regression analyses presented above, two dummy variables were created for group membership, namely Computer and Alpha.

Table 18 displays the results of the two separate regressions for the paper versus computer experiment. For all sub-groups, performing the composition item on computer had a positive effect on their scores. The size of the effect, however, was minimal for below average keyboarders. For fast keyboarders, however, the size of the effect increased sharply and was statistically significant.

For below average keyboarders, performing the composition on AlphaSmarts had a small negative affect on student performance. For fast keyboarders, however, performing the

composition item on AlphaSmart had a moderate and significantly positive affect on students' total scores.

Table 18: WPM Sub-group Regression Analyses for the Paper versus Computer Experiment

WPM<24 N=82					
R=.52	B	Std. Error	Beta	T	Sig.
Constant	4.33	1.52			
Mid-Year Grade	0.75	0.15	.50	5.15	<.001
Computer	0.84	0.63	.15	1.32	.19
Alpha	0.17	0.60	.03	0.28	.78
WPM>24 N=68					
R=.46	B	Std. Error	Beta	T	Sig.
Constant	7.26	2.15			
Mid-Year Grade	0.59	0.18	.37	3.27	.002
Computer	1.86	0.78	.31	2.41	.019
Alpha	1.18	0.81	.19	1.46	.15

Grade 8

The regression analyses presented above (Tables 10 and 11) indicate that mode of administration had a significant effect on students' performance for both the computer and eMate experiments. To test whether the effect of mode of administration varied for students with different levels of computer skill, students' WPM was used to form three groups. The first group contained students whose WPM was .5 standard deviations below the mean, or less than 23.8. The second group contained students whose WPM was between .5 standard deviations below the mean and .5 standard deviations above the mean, or between 23.8 and 32.4. The third group contained students whose WPM was .5 standard deviations above the mean or greater than 32.4. For each group, the composition total score was regressed on mid-term grades and group membership.

Table 97 displays the results of the three separate regressions for the paper versus computer experiment. For all sub-groups, performing the composition item on computer had a positive effect on their scores. Moreover, the size of this effect was nearly identical at all three levels of keyboarding speed.

Table 19: WPM Sub-group Regression Analyses for the Paper versus Computer Experiment

WPM<23.8 N=48					
R=.54	B	Std. Error	Beta	T	Sig.
Constant	-1.72	4.68			
Mid-Year Grade	0.17	0.06	.38	3.04	.004
Group	1.82	0.71	.32	2.56	.014
23.8>WPM<32.4 N=61					
R=.55	B	Std. Error	Beta	T	Sig.
Constant	-2.14	4.33			
Mid-Year Grade	0.19	0.05	.42	3.82	<.001
Group	1.46	0.05	.32	2.96	.004
WPM>32.4 N=31					
R=.45	B	Std. Error	Beta	T	Sig.
Constant	1.53	6.56			
Mid-Year Grade	0.16	.07	.37	2.22	.035
Group	0.17	.67	.29	1.76	.09

For students who wrote compositions on eMates, however, the size of the effect increased as keyboarding speed increased (Table 20). For students with slower and average keyboarding speed, the mode of administration effect was about the same size for eMates as the size on computers. For fast keyboarders, the mode of administration effect was noticeably larger.

Table 20: WPM Sub-group Regression Analyses for the Paper versus eMate Experiment

WPM<23.8 N=22					
R=.48	B	Std. Error	Beta	T	Sig.
Constant	-1.71	6.86			
Mid-Year Grade	0.17	0.08	.43	2.09	.049
Group	1.38	0.84	.33	1.63	.119
23.8>WPM<32.4 N=33					
R=.65	B	Std. Error	Beta	T	Sig.
Constant	-8.93	5.98			
Mid-Year Grade	0.26	0.07	.52	3.82	.001
Group	2.07	0.78	.37	2.67	.012
WPM>32.4 N=26					
R=.71	B	Std. Error	Beta	T	Sig.
Constant	-5.69	6.61			
Mid-Year Grade	0.22	0.07	.44	3.02	.006
Group	2.68	0.71	.55	3.77	.001

Grade 10 Sub-Group Analyses

The regression analyses presented above (Table 17) indicate that mode of administration had a significant effect on students' performance. To test whether the effect of mode of administration varied for students with different levels of computer skill, students' WPM was used to form three groups. The first group contained students whose WPM was .5 standard deviations below the mean, or less than 30.6. The second group contained students whose WPM was between .5 standard deviations below the mean and .5 standard deviations above the mean, or between 30.6 and 41.0. The third group contained students whose WPM was .5 standard deviations above the mean or greater than 41.0. For each group, the composition total score was regressed on mid-term grades and group membership.

Table 21 displays the results of the three separate regressions for the paper versus computer experiment. The largest effect was found for students with slower keyboarding skills. In this case, however, slow keyboarding skills ranged from 20 to 30 words per minute. In essence, no administration effect was found for average keyboarders, but a sizeable effect was found for fast keyboarders.

Table 21: WPM Sub-group Regression Analyses for the Paper versus Computer Experiment

WPM<30.6 N=43					
R=.52	B	Std. Error	Beta	T	Sig.
Constant	1.28	4.62			
Mid-Year Grade	0.13	0.06	.30	2.21	.033
Group	3.03	0.85	.48	3.58	.001
30.6>WPM<41.0 N=57					
R=.55	B	Std. Error	Beta	T	Sig.
Constant	-6.15	4.33			
Mid-Year Grade	0.25	0.05	.54	4.77	<.001
Group	0.38	0.66	.07	0.57	.568
WPM>41.0 N=42					
R=.31	B	Std. Error	Beta	T	Sig.
Constant	13.84	6.35			
Mid-Year Grade	0.002	0.08	.001	0.003	.99
Group	1.55	0.78	.31	1.98	.05

Examining Special Education and Mode of Administration

Many students receive special instructional and learning accommodations as result of learning challenges related to language arts. There is considerable debate, however, about whether students who receive special education accommodations in the classroom should also receive these accommodations during formal test administrations. One accommodation many special education students receive involves use of computers to take notes and complete classroom assignments. To examine whether the mode of administration effect differed for SPED students, the mode of administration effect was estimated separately for students receiving special education services for language arts and for those students who did not receive any special accommodations in the classroom. The effects for both groups of students are compared in Table 22.

As Table 22 indicates, the average keyboarding speed for third grade SPED students was considerably below the overall mean of 24. As reported in Russell's 1999 study, students whose keyboarding speed was below 20 words a minute generally do not benefit from performing open-ended items on computer. Given the low keyboarding speed of SPED students in third grade, it is difficult to determine whether the absence of a mode of administration effect is due to low keyboarding speed or is associated with the students' SPED status.

For students in grades eight and ten, Table 22 shows that the SPED students who performed the composition item on computer or with an eMate had lower mid-term grades than did the SPED students who wrote their responses on paper. For both grades eight and ten, the differences between the two groups mid-term grades was statistically significant. To control for these differences in prior achievement, a series of multiple regressions was performed within all three grade levels. Within each grade level, total score was regressed on mid-term grades and on group membership (paper = 0, computer/eMate = 1). Within each grade, separate regressions were performed for SPED students and for Non-SPED students. Table 23 presents the regression coefficients and standardized effects within each grade level.

As Table 23 indicates, the effect size for SPED students in grade four was negative. Recall, however, that these students were generally poor keyboarders and that poor keyboarders generally do not benefit from performing written tests on a computer.

In grade eight, the mode of administration effect after controlling for differences in mid-term grades was about 1.5 times larger for SPED students than for Non-SPED students. This pattern, however, does not hold in grade ten.

Table 22: Mode of Administration Effect for SPED and Non-SPED Students

	SPED Means			Non-SPED Means		
	Paper	Comp.	Effect Size	Paper	Comp.	Effect Size
Grade 4						
Mid-Term Grade	9.9	9.0	-.58	10.9	11.3	.23
WPM	15.3	14.5	-.12	23.6	25.7	.03
Total Score	11.9	9.6	-.80	13.0	15.2	.87*
N	9	7		40	43	
Grade 8						
Mid-Term Grade	90.5	87.3	-.70*	86.3	88.0	.27
WPM	31.9	27.8	-.44	26.5	28.6	.22
Total Score	14.3	15.8	.58	13.7	15.7	.79*
N	26	22		101	79	
Grade 10						
Mid-Term Grade	78.5	72.7	-1.26*	82.4	84.0	.26
WPM	26.8	30.1	.39	36.5	37.3	.09
Total Score	12.4	11.4	-.41	13.4	15.3	.61*
N	10	9		64	62	

* Statistically significant at the .05 level

Table 23: Effect Controlling for Mid-Term Grade

	B	SE	Group Beta	T	Sig.
Grade 4					
Non-SPED	1.89	0.56	.31	3.38	.001
SPED	-1.17	1.04	-.23	1.13	.28
Grade 8					
Non-SPED	1.65	0.33	.31	5.00	<.001
SPED	2.51	.060	.48	4.18	<.001
Grade 10					
Non-SPED	1.66	.49	.27	3.39	.001
SPED	-1.10	0.99	-.29	1.11	.28

Discussion

The series of experiments described here extends the work of Russell (1999) and Russell and Haney (1997) in two important ways. In addition to examining the mode of administration effect in grade eight, this study examines the effect in grades four and ten. In many state testing programs as well as the National Assessment of Educational Progress and international studies such as the Third International Mathematics and Science Study and PIRLS, these three grade levels are commonly tested. Thus, it is important to understand the extent to which mode of administration affects the performance of students in these commonly tested grade levels. Second, this series of studies introduces a third mode of administration that provides schools and testing programs with a cheaper word processing option, namely AlphaSmarts and eMates.

As in the two previous studies, this series of studies found that students who wrote their compositions on computer produced longer responses that received higher scores. These effects were larger in grade eight than in grade ten, but were statistically and practically significant at all grade levels. In addition, substituting an eMate for a desktop computer also had a positive effect on students' performance in grade eight.

Across the three grade levels, the relationship between keyboarding speed and the mode of administration effect was inconsistent. As Russell (1999) found, the fourth grade study presented above indicates that students need to have sufficient keyboarding speed before the mode of administration effect becomes meaningful. In grade eight, however, this pattern did not emerge. Although the effect was largest for fast keyboarders who produced their composition on an eMate, the effect was positive and of practical significance at all levels of keyboarding speed and in nearly all experiments. One reason why this finding differs from that reported by Russell (1999) may stem from the relatively high level of keyboarding speed for all students included in this series of experiments. Whereas Russell (1999) reported an average keyboarding speed of about 15 words per minute, average speed for the grade eight students included in this study was nearly twice as fast. Moreover, the average speed of students included in the low keyboarding level of this study exceeded the cut point for the high keyboarding level in Russell's (1999) study. It appears that once students achieve keyboarding speed of roughly 20 wpm, the mode of administration effect becomes meaningful. The size of the effect, however, does not increase further as students keyboarding speed increases.

The mode of administration effect, however, may differ for students who receive special education services for language arts. For students in grade eight, the mode of administration effect was about one and a half times larger for SPED students than for non-SPED students. This finding did not hold in grade ten. Nonetheless, this issue warrants further exploration.

Limitations

This series of studies focused on students in grades four, eight and ten attending school within a single district. This district tends to perform well above the state average on standardized and state level tests. In the studies presented here, very few students performed at low levels on the composition item. As a result, it was not possible to examine the mode of administration effect across the full range of performance levels.

Similarly, students included in this study were generally accustomed to working with computers. The relatively high level of keyboarding speed complicated efforts to examine the

mode of administration effect at low levels of keyboarding speed. Additionally, students' familiarity with computers prevented an examination of the mode of administration effect for students who are not accustomed to working with computers. Despite this high level of access to computers within Wellesley's schools, it should be noted that seventy-five out of approximately three hundred districts across the state have a better student to computer ratio than does Wellesley.

Although this study attempted to compare the mode of administration effect for SPED students with Non-SPED students, the moderate sample size combined with the relatively small number of SPED students frustrated these analyses. Moreover, the small number of SPED students prevented examining the interaction between keyboarding speed, special education and the mode of administration effect.

Since students' scores on the composition item did not "count" toward students' grades and did not become part of students' records, many students in grade ten may not have put forth their best efforts. This was particularly true for the computer group who was proctored by a person who was not a teacher in the high school. Students performing the test on paper, however, were proctored by their English teacher. Although there were no severe behavioral problems reported, most students in the computer group were reported to have finished their compositions within the first fifteen minutes of the second writing period. Students in the paper groups, however, worked for nearly the entire period. Moreover, four responses generated on computer failed to address the question posed, whereas all of the responses composed on paper were on topic. For these reasons, data collected in grade ten may not accurately reflect students' actual levels of performance.

Implications

This series of studies provides further evidence that students accustomed to writing on computers perform significantly better when open-ended tests are administered on a computer. For the MCAS Language Arts test, this improved performance translates into approximately two points on the composition item. But in addition to the composition item, the MCAS Language Arts test contains four short answer open-ended items each worth four points. Assuming that the effect found for the composition item and that the effect reported on the shorter open-ended items by Russell (1999) holds across all Language Arts open-ended items, students accustomed to writing with computers may perform better by an additional two points on these open-ended items. Across all short and extended open-ended MCAS items, the mode of administration effect may result in an increase of four raw score points if students were allowed to compose responses on a computer. An increase of four raw score points translates into between four and eight point scale score points, depending upon where on the scale a student's score resides. This score increase may be even larger for fast keyboarders in grade four. Clearly, as state level testing programs such as MCAS begin and/or continue to use test scores to make critical decisions about graduation and promotion, steps should be taken that allow students who are accustomed to working on computers to perform at their highest level.

School Level Effect

Within Wellesley, eliminating the mode of administration effect for both the composition item and the four shorter open-ended items would have a dramatic impact on district level results. As figure 5 indicates, based on last years (1999) MCAS results, 19% of the fourth

graders classified as "Needs Improvement" would move up to the "Proficient" performance level. An additional 5% of students who were classified as "Proficient" would be deemed "Advanced." Similarly, figure 6 shows that in grade eight, four percent of students would move from the "Needs Improvement" category to the "Proficient" category and that 13% more students would be deemed "Advanced." In grade 10, allowing students to perform written items on computer would move 14% more students into the "Advanced" category and approximately 7% more students into the "Proficient" category (Figure 7). As Figure 8 displays, within one elementary school (Bates), the percentage of students performing at or above the "Proficient" level would nearly double from 39% to 67%.

Figure 5: Mode of Administration Effect on Grade 4 1999 MCAS Results

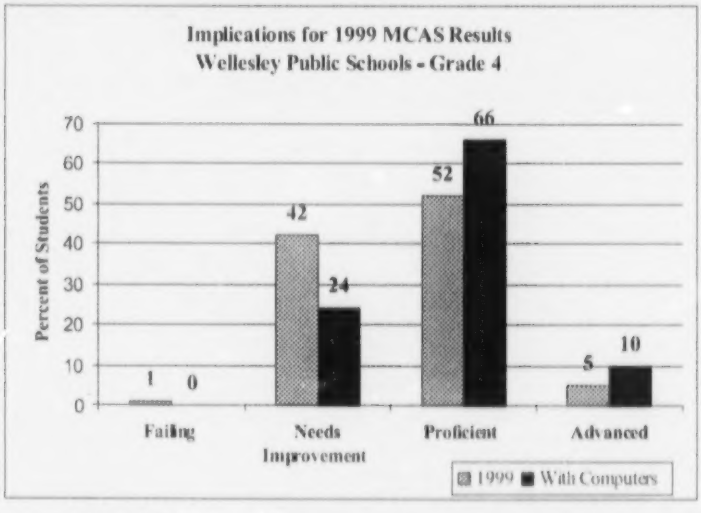


Figure 6: Mode of Administration Effect on Grade 8 1999 MCAS Results

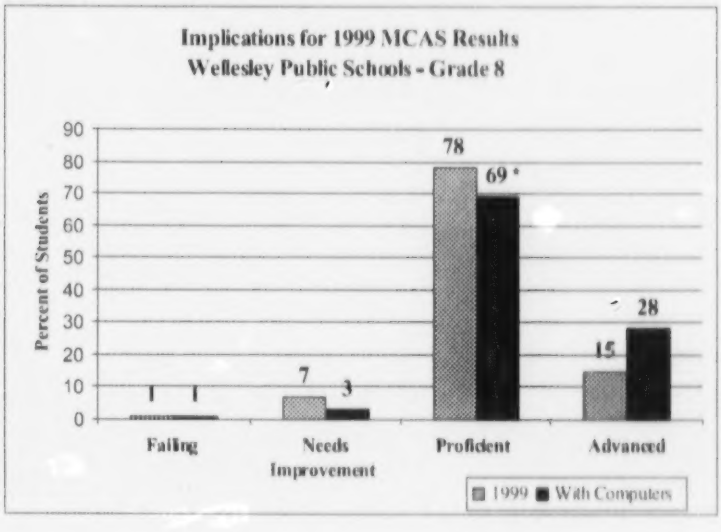


Figure 7: Mode of Administration Effect on Grade 10 1999 MCAS Results

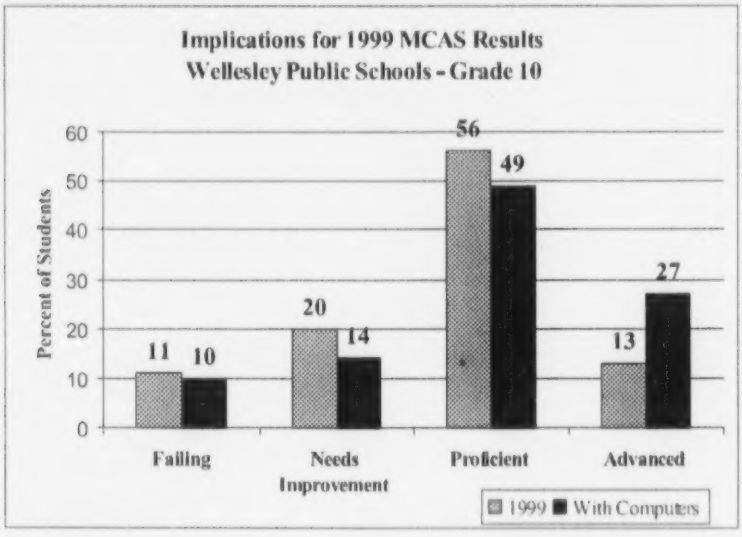
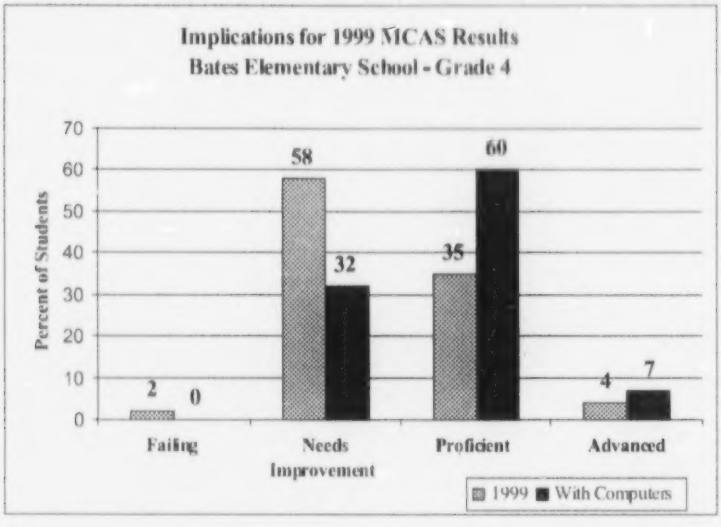


Figure 8: Mode of Administration Effect on Bates Elementary School 1999 MCAS Results



Policy Options

One solution state testing programs might adopt to reduce the mode of administration effect is to allow students to select the mode in which open-ended responses are composed. For the past decade, the Province of Alberta has employed this strategy for its graduation testing program. Over the past five years, the province has seen the percentage of students opting to perform the English, Social Studies, Biology and French tests on computer increase from 6.7% in 1996 to 24.5% in 2000. Within high schools, the percentage of students opting to perform the test on a computer ranges from 0 to 80% (Sakyi, 2000).

Although this approach adds to the complexity to test administration procedures and raises security concerns in schools that do not have sufficient word processors for all students to perform a test simultaneously, an additional complication arises during the scoring process. Although there has not been a large amount of research on the extent to which computer printing versus hand-writing affects ratings of written work, Powers et. al. (1994) report that significant effects can occur. Surprisingly, Powers et. al. found that computer printed responses produced by adults tended to receive lower scores than the same responses produced by hand.

To explore the extent to which hand-written versus computer printed text effects raters' scores, sixty responses from each grade level originally produced by hand were presented to raters in three ways: handwritten, single spaced twelve point computer text, and double-spaced fourteen point computer text. A spiral design in which raters scored twenty of each presentation format was used to reduce rater bias. As in the series of mode of administration studies presented above, all responses were double-scored using the MCAS rubrics and scores from the two raters were summed.

Table 24 presents the mean score for each mode of presentation by grade levels. In all grade levels, the mean score for responses presented in handwriting was higher than mean scores for responses presented as computer text. In grade eight, the presentation effect resulted in more than a two point difference between scores awarded to handwritten responses versus computer printed responses. In grade eight, there was little difference between scores awarded to single spaced and double spaced responses. In grade ten, however, double-spaced responses tended to be awarded lower scores than single-spaced responses.

Table 24: Descriptive Statistics for Mode of Presentation

	Handwritten		Single Spaced		Double Spaced	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Grade 4						
Topic Development	8.3	1.71	7.0	1.99	7.4	2.06
English Standards	6.5	1.11	5.8	1.26	5.7	1.38
Total Score	14.8	2.57	12.8	3.04	13	3.29
Grade 8						
Topic Development	9.2	1.69	7.9	1.67	7.9	1.78
English Standards	6.9	0.97	6.0	1.07	6.1	1.17
Total Score	16.1	2.51	13.9	2.50	14.0	2.71
Grade 10						
Topic Development	8.6	2.33	7.7	2.32	6.9	2.22
English Standards	6.8	1.27	6.2	1.33	6.0	1.37
Total Score	15.4	3.41	13.9	3.24	12.9	3.23

To examine whether the total score mean differences were statistically significant, a one way analysis of variance was performed within each grade level. As Table 25 indicates, the differences were significant within each grade level. To compare total score means for each mode of presentation, Scheffe post-hoc comparisons were performed. Table 26 indicates that for grades four, eight and ten, differences between both single-spaced and handwritten responses and between double-spaced and handwritten responses were statistically significant. Differences between single and double-spaced responses were not statistically significant.

Table 25: Mode of Presentation Total Score ANOVA

	Mean Square			
	Between	Within	F	Sig.
Grade 4 Total Score	59.5	8.9	6.7	.002
Grade 8 Total Score	92.7	6.6	14.0	<.001
Grade 10 Total Score	92.9	10.8	8.6	<.001

Table 26: Scheffe Multiple Comparisons for Mode of Presentation Total Scores

	Single vs Handwritten			Double vs. Handwritten			Single vs. Double		
	Mean Diff.	Std. Error	Sig.	Mean Diff.	Std. Error	Sig.	Mean Diff.	Std. Error	Sig.
Grade 4	1.9	.58	.004	1.69	.58	.017	.29	.58	.89
Grade 8	2.22	0.47	<.001	2.08	0.47	<.001	0.13	.47	.96
Grade 10	1.52	0.60	.04	2.47	0.60	<.001	0.95	.60	.29

Given the mode of presentation effect, it appears that simply offering students the choice of working on computer or on paper may still result in under-estimates of the performance of students working on computer (or over-estimates of students working on writing).

In the short-term, then, it appears the best policy option is to recognize the limitations of information gleaned from open-ended language arts items. Before scores from open-ended items are used to make high-stakes decisions about students and their schools, further research into the mode of administration effect and into strategies to reduce the mode of presentation effect are clearly needed. Furthermore, future studies in this area should be performed during actual test administrations and across multiple districts.

Without question, both computer technologies and performance testing can help improve the quality of education. However, until students' can take tests in the same medium in which they generally work and learn, we must recognize that the scores from high-stakes state tests do not provide authentic measures of some students' capabilities. Reliance on paper and pencil written test scores to measure or judge student and/or school performance will mischaracterize the achievement of students' accustomed to working on computers.

Administration Guidelines

The mode of administration effects found in this series of studies provide further evidence that students who are accustomed to writing on computers perform significantly better on open-ended items that are administered on computer. As increasing numbers of students become accustomed to writing on a computer, this mode of administration effect will impact the performance of a larger percentage of students. As state-level testing programs, such as MCAS, mature, it is vital that they begin developing plans to administer written portions of exams on a computer.

As reported above, the province of Alberta has provided schools and students the option of completing written portions of their English, Social Studies, Biology and French graduation exams on paper or on computer since 1996. In their test administration guidelines, Alberta justifies providing this option as follows:

"Producing written work on computer is common in contemporary working and learning environments. Students who have been taught to compose on a computer, and who normally produce their written work in this way, would be disadvantaged if they were compelled to hand-write extended assignments." (Alberta Learning, 2000, pg. 55)

Through our experiences administering the MCAS composition items on computer to a limited sample of students across a single district combined with documents developed by the province of Alberta, we strongly encourage state-level testing programs to consider adopting the administrative procedures described below. As web-based test administration and connectivity in schools improves, some of these guidelines may become obsolete. However, in the short-term these guidelines should enable schools to provide students the option of performing open-ended items on computer with a minimal of additional effort while maintaining test security.

Overall Guiding Principle

A school principal should provide students the choice of performing open ended Language Arts items using paper and pencil or using a computer. All students should be extended this option.

General Guidelines

- Principal of a school will decide whether or not to provide students opportunity to write their exams on school owned computers and software
- Principal should respond positively if students normally write on computer, students know how to use hardware and software at school, technical expertise is available before, during, and after exam, and security, fairness, and confidentiality in no way will be compromised
- Once the principal has decided to provide this option, the decision to work on paper or on computer will be made by each individual student
- If the school has more students who require this provision than it has computers, principal will submit a proposal for meeting needs to the Department of Education
- First time a school provides students this option, they will perform a practice session to familiarize themselves and students with procedures before the actual exam
- Make sure all computers tested in good working order, wordprocessor installed, correct print drivers selected, and that computers print to selected printer
- On the top of each students document, a header must be placed containing students last name, first name, student identification number, and page number
- On first line of each document student's name must be typed
- Each computer's word processing program is opened and formatted in advance of exam writing

During the exam

- A technician or other person familiar with the computers and software used must be on hand at all times in each testing room to address any technical glitches that occur

After the exam

- All student responses must be printed out in their entirety and stapled directly into each student's test booklet
- Once the exam proctor is assured that he/she has a copy of the final writing document from all students, technical personnel must make sure no work remains anywhere on the school district's computer system

Exam Proctor's Role

- Students should have access to routine word processing tools integrated into the software such as spell checkers and thesaurus but no grammar checker allowed.
- Internet access and online resources not available
- Frequent and quiet printing and distribution of hard copies must be possible
- Students may revert to pencil and paper at any time
- Students are encouraged to save on their computer on a regular basis or make use of automatic save feature
- Students must print their work in a pre-assigned double-spaced 12 point font
- Final printing may extend beyond officially scheduled exam time as needed

Potential items to work out by Department of Education ahead of time

- More students desire to use computers than you have computers available. Should the exam be administered twice during the same day? Or be offered a different day using a different writing prompt?
- Should printed copies of both the first and final drafts be collected?
- Is there a way to submit student responses electronically via the World Wide Web?

Further Research

Several small experiments conducted within Massachusetts have demonstrated that students who are accustomed to writing on computers perform better on tests when they are allowed to use a computer to compose written responses. Together, these studies also suggest that students must be able to keyboard moderately well (about 20-24 words per minute) before the mode of administration effect becomes significant. Moreover, at low levels of keyboarding speed, the effect seems to be negative. The most recent set of studies also provide some evidence that the mode of administration effect may differ for SPED and non-SPED students.

In all of these studies, however, moderate sample sizes and relatively homogenous populations (that is within a single school and/or district) limited an examination of the relationships between the mode of administration effect and issues such as keyboarding speed, SPED status, regular use of computers in school versus at home, and access to spell-checking.

In addition, the studies performed to date have not allowed us to test the administrative guidelines outlined above or to explore alternative methods of administering and collecting responses (e.g., via the world wide web).

Finally, none of the studies conducted to date have occurred during actual testing time. As a result, it is possible that some students have not put forth their best efforts. Although it has been assumed that motivation and effort would impact equally the performance of students working on paper or on computer, this assumption has not been tested.

Despite the limitations of the studies performed to date, the mode of administration effects reported here and in previous studies (Russell, 1999; Russell and Haney, 1997) highlight a complicated challenge testing programs must overcome as they continue to use tests containing open-ended items to make inferences about student and school achievement. Clearly, as high-stakes decisions are linked to test scores based in part or entirely on students' performance on open-ended items, further research into the mode of administration effect, presentation effects and strategies for reducing both effects is needed.

References

Alberta Learning. (2000). Directions for Administration, Administrators Manual, Diploma Examination Program.

Arnold, V., Joyner, R. L., Schmidt, B. J., & White, C. D. (1997). Establishing electronic keyboarding speed and accuracy timed-writing standards for postsecondary students. *Business Education Forum*, 51(3), p. 33-38.

Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research*, 63(1), 69-93.

Beaton, A. E. & Zwick, R. (1990). *The Effect of Changes in the National Assessment: Disentangling the NAEP 1985-86 Reading Anomaly*. Princeton, NJ: Educational Testing Service, ETS.

Becker, H. J. (1999). *Internet Use by Teachers: Conditions of Professional Use and Teacher-Directed Student Use*. Irvine, CA: Center for Research on Information Technology and Organizations.

Bunderson, C. V., Inouye, D. K. & Olsen, J. B. (1989). The four generations of computerized educational measurement. In Linn, R. L., *Educational Measurement* (3rd ed.), Washington, D.C.: American Council on Education, pp. 367-407.

Burstein, J., Kaplan, R., Wolff, S., & Lu, C. (1997). *Using lexical semantic techniques to classify free-responses*. A report issued by Educational Testing Service. Princeton, NJ.

Cizek, G. J. (1991). The effect of altering the position of options in a multiple-choice Examination. Paper presented at NCME, April 1991. (ERIC)

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Hillsadale, NJ: Erlbaum.

Council of Chief State School Officers (1998). *Key State Education Policies on K-12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance*. Washington, DC.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

Daiute, C. (1986). Physical and cognitive factors in revising: insights from studies with computers. *Research in the Teaching of English*, 20 (May), p. 141-59.

Educational Testing Service. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics*. Princeton, NJ: Policy Information Center Research Division Educational Testing Service.

Etchison, C. (1989). Word processing: A helpful tool for basic writers. *Computers and Composition*, 6(2), 33-43.

Fabos, B. & Young, M. (1999). Telecommunications in the classroom: Rhetoric versus reality. *Review of Educational Research*. 69:3 217-259.

Glass, G. & Hopkins, K. (1984). *Statistical Methods in Education and Psychology*. Boston, MA: Allyn and Bacon.

Glennan, T. K., & Melmed, A. (1996). *Fostering the use of educational technology: Elements of a national strategy*. Santa Monica, CA: RAND.

Green, B. F., Jr. (1970). Comments on tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing and guidance. New York: Harper and Row.

Grejda, G. F. (1992). Effects of word processing on sixth graders' holistic writing and revision. *Journal of Educational Research*, 85(3), 144-149.

Haas, C. & Hayes, J. R. (1986a). Pen and paper versus the machine: Writers composing in hard-copy and computer conditions (CDC Technical Report No. 16). Pittsburgh, PA: Carnegie-Mellon University, Communication Design Center. (ERIC ED 265 563).

Haney, W., Madaus, G., & Lyons, R. (1993). *The Fractured Marketplace for Standardized Testing*. Boston, MA: Kluwer Academic Publishers.

Hannafin, M. J. & Dalton, D. W. (1987). The effects of word processing on written composition. *The Journal of Educational Research*, 80 (July/Aug.) p. 338-42.

Hawisher, G. E. & Fortune, R. (1989). Word processing and the basic writer. *Collegiate Microcomputer*, 7(3), 275-287.

Hawisher, G. E. (1986, April). The effect of word processing on the revision strategies of college students. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC ED 268 546)

Holmes, R. (1999). A gender bias in the MCAS? *MetroWest Town Online*. http://www.townonline.com/metrowest/archive/022499/.

Kerchner, L. B. & Kistinger, B. J. (1984). Language processing/word processing: Written expression, computers, and learning disabled students. *Learning Disability Quarterly*, 7(4), 329-335.

Kurth, R. J. (1987). Using word processing to enhance revision strategies during student writing activities. *Educational Technology*, 27(1), 13-19.

MacArthur C. & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing and dictation. *Journal of Special Education*, 21(3), 22-42.

MacArthur, C. A. (1988). The impact of computers on the writing process. *Exceptional Children*, 54(6), 536-542.

Market Data Retrieval. (1999). *Technology in Education 1999*. (A report issued by Market Data Retrieval). Shelton, CN: Market Data Retrieval.

Massachusetts Department of Education. (1999). 1998 MCAS Technical Report Summary. Malden, MA.

Mead, A. D. & Drasgow, (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114:3, 449-58.

Mourant, R. R., Lakshmanan, R. & Chantadisai, R. (1981). Visual Fatigue and Cathode Ray Rube Display Terminals. *Human Factors*, 23(5), 529-540.

Nichols, L. M. (1996). Pencil and paper versus word processing: a comparative study of creative writing in the elementary school. *Journal of Research on Computing in Education*, 29(2), 159-166.

Oppenheimer, T. (1997). The computer delusion. *The Atlantic Monthly*. 280(1), 45-62.

Owston, R. D. (1991). Effects of word processing on student writing in a high-computer-access environment (Technical Report 91-3). North York, Ontario: York University, Centre for the Study of Computers in Education.

Phoenix, J. & Hannan, E. (1984). Word processing in the grade 1 classroom. *Language Arts*, 61(8), 804-812.

Powers, D., Fowles, M, Farnum, M, & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31(3), 220-233.

Powers, D., Fowles, M, Farnum, M, & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31(3), 220-233.

Robinson-Stavely, K. & Cooper, J. (1990). The use of computers for writing: Effects on an English composition class. *Journal of Educational Computing Research*, 6(1), 41-48.

Robison, J. W., Erickson, L. W., Crawford, T.J. & Ownby, A. C. (1979). *Typewriting: Learning and Instruction*. Cincinnati: South-Western Publishing Company.

Russell, M. & Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, 5(3), http://olam.ed.asu.edu/epaa/v5n3.html.

Russell, M. & Haney, W. (2000). Bridging the Gap Between Testing and Technology in Schools. *Education Policy Analysis Archives*, 8(19),http://epaa.asu.edu/epaa/v8n19.html.

Russell, M. (1999). Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Educational Policy Analysis Archives*, 7(20).

Russon, A. R. and Wanous, S. J. (1973). *Philosophy and Psychology of Teaching Typewriting*. (2nd ed.) Cincinnati: South-Western Publishing Company.

Sacks, P. (1999). *Mind Games: How our mental testing culture sustain the privileged, punishes the poor and standardizes the American mind*. Reading MA: Perseus Books.

Sakyi, A. (2000). Word Processing in Humanities Diploma Examinations. Personal correspondence from the Alberta Learning, Student Evaluation Branch.

Sitko, M.C. & Crealock, C. M. (1986, June). A longitudinal study of the efficacy of computer technology for improving the writing skills of mildly handicapped adolescents. Paper presented at at the Invitational Research Symposium on Special Education Technology, Washington, DC.

Snyder, T. D. & Hoffman, C. (1990). *Digest of Education Statistics*. Washington, DC: U. S. Department of Education.

Snyder, T. D. & Hoffman, C. (1994). *Digest of Education Statistics*. Washington, DC: U. W. Department of Education.

Vacc, N. N. (1987). Word processor versus handwriting: A comparative study of writing samples produced by mildly mentally handicapped students. *Exceptional Children*, 54(2), 156-165.

West, L. J. (1968) The vocabulary of instructional materials for typing and stenographic training - research findings and implications. *Delta Pi Epsilon Journal*, 10(3), 13-125.

West, L. J. (1983). *Acquisition of Typewriting Skills*, (2nd ed.), Indianapolis: The Bobbs-Merrill Company, Inc.

Williamson, M. L. & Pence, P. (1989). Word processing and student writers. In B. K. Briten & S. M. Glynn (Eds.), *Computer Writing Environments: Theory, Research, and Design* (pp. 96-127). Hillsdale, NJ: Lawrence Erlbaum & Associates.

Wolf, F. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Sage University series on quatitative applications in the social sciences, series no. 07-059. Newbury Park, CA: Sage.

Zandvliet, D. & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, 29(4), 423-438.

About the National Board on Educational Testing and Public Policy

Created as an independent monitoring system for assessment in America, the National Board on Educational Testing and Public Policy is located in the Peter S. and Carolyn A. Lynch School of Education at Boston College. The National Board provides research-based test information for policy decision making, with special attention to groups historically underserved by the educational systems of our country. Specifically, the National Board

- Monitors testing programs, policies, and products
- Evaluates the benefits and costs of testing programs in operation
- Assesses the extent to which professional standards for test development and use are met in practice

This National Board publication series is supported by a grant from the Ford Foundation.

The National Board on Educational Testing and Public Policy
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Telephone: (617)552-4521 • Fax: (617)552-8419
Email: nbetpp@bc.edu

Visit our website at **nbetpp.bc.edu** for more articles, the latest educational news, and for more information about NBETPP.

NBETPP

The Board of Directors

Peter Lynch
Vice Chairman
Fidelity Management and Research

Paul LeMahieu
Superintendent of Education
State of Hawaii

Donald Stewart
President and CEO
The Chicago Community Trust

Antonia Hernandez
President and General Council
Mexican American Legal Defense and Educational Fund

Faith Smith
President
Native American Educational Services

BOSTON COLLEGE